Published online 28 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D915–D919***
doi:10.1093/nar/gks1245

H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery

Jun-ichi Takeda[1], Chisato Yamasaki[1], Katsuhiko Murakami[1], Yoko Nagai[1], Miho Sera[1], Yuichiro Hara[1], Nobuo Obi[1], Takuya Habara[1], Takashi Gojobori[1,2] and Tadashi Imanishi[1,*]

[1]Integrated Database and Systems Biology Team, Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, Aomi 2-4-7, Koto-ku, Tokyo 135-0064, Japan and [2]Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Received September 14, 2012; Revised and Accepted November 2, 2012

ABSTRACT

H-InvDB (http://www.h-invitational.jp/) is a comprehensive human gene database started in 2004. In the latest version, H-InvDB 8.0, a total of 244 709 human complementary DNA was mapped onto the hg19 reference genome and 43 829 gene loci, including nonprotein-coding ones, were identified. Of these loci, 35 631 were identified as potential protein-coding genes, and 22 898 of these were identical to known genes. In our analysis, 19 309 annotated genes were specific to H-InvDB and not found in RefSeq and Ensembl. In fact, 233 genes of the 19 309 turned out to have protein functions in this version of H-InvDB; they were annotated as unknown protein functions in the previous version. Furthermore, 11 genes were identified as known Mendelian disorder genes. It is advantageous that many biologically functional genes are hidden in the H-InvDB unique genes. As large-scale proteomic projects have been conducted to elucidate the functions of all human proteins, we have enhanced the proteomic information with an advanced protein view and new subdatabase of protein complexes (Protein Complex Database with quality index). We propose that H-InvDB is an important resource for finding novel candidate targets for medical care and drug development.

INTRODUCTION

Along with the sequencing of the first human reference genome (1), several lines of human transcriptome study using a large number of validated human transcripts were carried out. As full-length complementary DNA (cDNA) is the ideal resource for the study, our consortium aimed to collect human full-length cDNA sequenced by four projects: Full-Length cDNA Japan (FLJ) (2), Human Unidentified Gene-Encoded Large Proteins (HUGE) (3), Mammalian Gene Collection (MGC) (4) and Munich Information Centre for Protein Sequences (MIPS) (5). These projects were conducted at five institutions: New Energy and Industrial Technology Development Organization (NEDO), Kazusa DNA Research Institute (KDRI), the National Institutes of Health (NIH, USA), German Research Centre for Environment and Health (GSF) and Chinese National Human Genome Centre (CHGC) (6). Our consortium then held an international workshop called Human Full-Length cDNA Annotation Invitational (H-Invitational or H-Inv) to manually annotate the registered human full-length cDNA sequences on our annotation system by expert scientists and annotators (7). To release the annotation results, the first H-InvDB was constructed in 2004, and as of the third version in 2006, H-InvDB was extended to include all published human cDNA in addition to H-Inv human full-length cDNA (8).

At present, H-InvDB has been developed as not only a human transcriptome database but also one of the largest integrative human omics databases available to human gene researchers in various biological fields. One of the features of H-InvDB is that all published human cDNA sequences were annotated by a rigorous annotation pipeline confirmed at H-Invitational (7,9). For example, we examine sequence quality, sequence identity with the human reference genome sequence, sequence orientation (some cDNA sequences are registered in reverse direction), chimeric or truncated cDNAs and possible contamination from other species. Thus, most artifacts were removed and misannotations were expected to be few. H-InvDB also contains several specific H-Inv sub and satellite databases based on the annotation of H-Inv human transcripts (Figure 1). Databases involving gene expression (H-ANGEL) (10), molecular evolution (Evola) (11),

[*]To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: t.imanishi@aist.go.jp

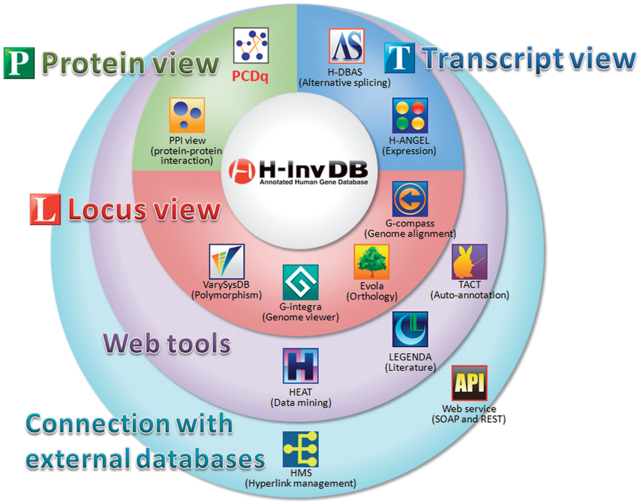


Figure 1. A schematic diagram of H-InvDB as a central hub for human omics study. Each content is described shortly in the Quick guide page (http://h-invitational.jp/hinv/ahg-db/tools.jsp).

genetic polymorphism (VarySysDB) (12) and alternative splicing (H-DBAS) (13) have been developed. Thus, users can find objective human annotation information in diverse combinations by using the search system of H-InvDB. In addition to these databases, H-InvDB is also connected with external databases by the web service application program interfaces (APIs) and Hyperlink Management System (HMS) (14). On these accounts, H-InvDB is a reliable and useful database for omics studies.

CHARACTERISTICS OF H-InvDB RELEASE 8.0

Update information

In the latest version of H-InvDB 8.0, 244 709 human transcript sequences extracted from DDBJ (15) were freshly mapped on the assembled reference genome UCSC hg19 (16). Clustering the transcripts revealed 43 829 gene loci called H-Inv clusters (HIXs) (Table 1). Among these 43 829 genes, 35 631 were predicted as potential protein-coding genes. This number is much larger than the number of nonredundant protein entries in UniProtKB/SwissProt (17), which is a literature-based, human curated database of known proteins, because H-InvDB contains both known and predicted proteins from human transcripts. We classified them into seven protein categories according to the strength of protein evidence (7) and found that 22 898 genes were predicted to have at least one protein functional motif (Categories I–III) (Table 2).

Including all these protein categories, all H-Inv transcripts (HITs) were annotated with various sequence features, such as gene structures, alternative splicing variants, noncoding functional RNA, protein functions, functional domains, subcellular localizations, metabolic pathways, protein 3D structure, genetic polymorphisms (single-nucleotide polymorphism, indels and microsatellite repeats), association with diseases, gene expression profiling, molecular evolutionary features, protein–protein interactions (PPIs) and gene families/groups.

Table 1. Statistics of H-InvDB 8.0

Number of gene clusters (HIX)	Number of transcripts (HIT)	Number of proteins (HIP)
43 829	244 709	147 684

Table 2. Statistics of representative HIPs

Category	Definition	Number of representative HITs
I	Identical to known human protein (≥98% identity and 100% coverage)	16 128
II	Similar to known protein (≥50% identity and ≥50% coverage)	5872
III	InterPro domain containing protein	898
IV	Conserved hypothetical protein	1705
V	Hypothetical protein	5268
VI	Hypothetical short protein (20–79 amino acids)	5068
VII	Pseudogene candidates	692
Total		35 631

These annotations were assigned to not only H-InvDB but also the corresponding specific H-Inv sub and satellite databases in detail. These annotations are also used as search items in the H-InvDB Navi system (8) for compound retrieval. Among the H-Inv satellite databases, H-InvDB Enrichment Analysis Tool (HEAT) (8) was considerably upgraded. HEAT is a tool for gene-set enrichment analysis based on various annotation in H-InvDB, such as InterPro (18), GO (19), KEGG pathway (20), SCOP (21), subcellular localization, chromosomal band, gene family and tissue specific expression in H-ANGEL (10). It searches for H-InvDB annotations that are significantly enriched in a user-defined gene sets as compared with the entire H-InvDB representative protein-coding transcripts. We newly added promoter motifs of all human genes based on JASPAR (22) and PPIs in the HEAT system. This enabled us to conduct extensive data mining with the HEAT system.

Advantages of H-InvDB

We compared 43 829 H-Inv genes with RefSeq (23) and Ensembl (24) genes to enumerate the numbers of unique and overlapping entries. Although the numbers of H-Inv unique genes were similar to those of the Ensembl unique ones (19 309 and 19 063, respectively) (Figure 2A), H-InvDB uses only rigorously annotated human cDNA sequences, including those of experimentally validated full-length cDNA (7). These characteristics suggest that our uniquely annotated genes were likely to be biologically functional. To investigate the evidence for protein coding of H-Inv unique genes, we also compared frequencies of the genes in protein categories between H-InvDB and the consensus coding sequence (CCDS) (25) (Figure 2B). The result indicated that unknown functional proteins (Categories V and VI) and nonprotein-coding sequences were frequent in H-Inv unique genes. As described earlier, these H-Inv unique unknown proteins were completely

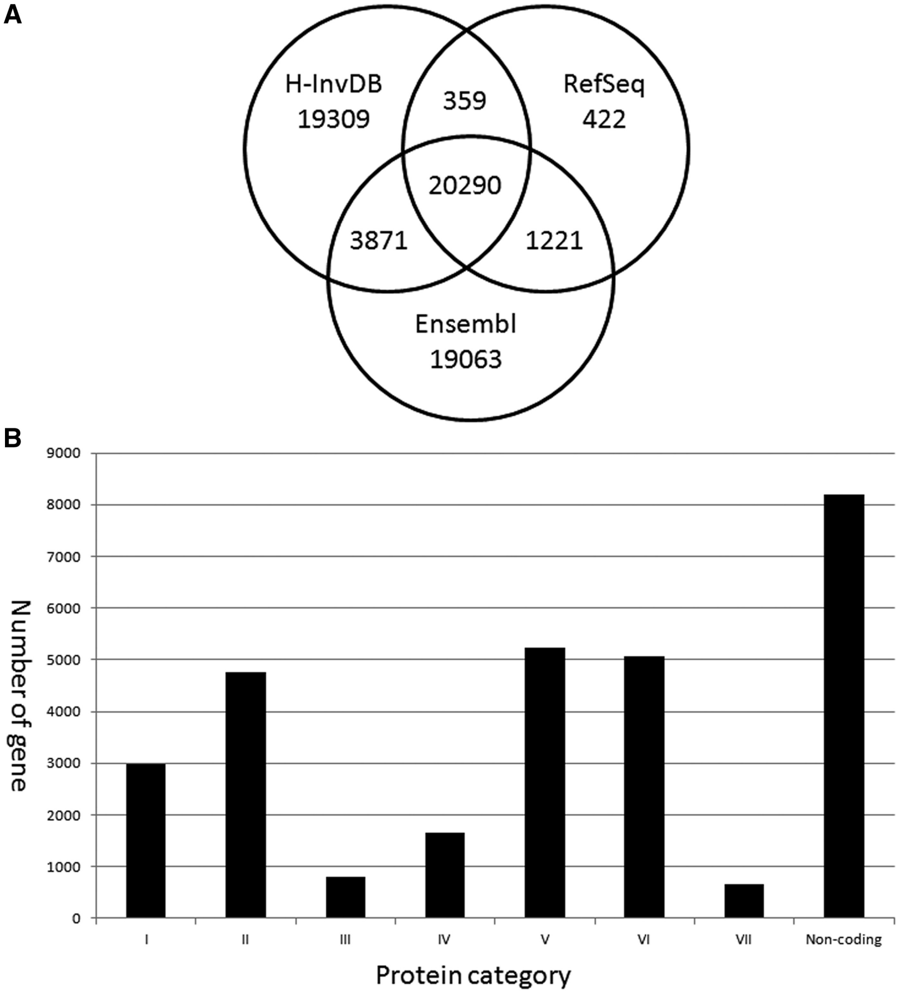


Figure 2. Comparison of gene numbers between H-InvDB and other databases. **(A)** The Venn diagram represents the numbers of unique and overlapping genes among H-InvDB, RefSeq and Ensembl. **(B)** The bar graph represents the numbers of H-Inv unique genes when compared with CCDS genes. The roman numerals indicate protein categories shown in Table 2.

transcribed as they can indeed have some functions. In fact, 233 genes, which have been classified as hypothetical proteins (Categories V and VI) in the previous version of H-InvDB 6.2, turned out to be functional proteins (Categories I–III) in the latest version of H-InvDB 8.0, because they were found in Online Mendelian Inheritance in Man (OMIM) (26) (Table 3). Among them, 11 Category I genes were suggested to associate with Mendelian disorders based on OMIM (Supplementary Table S1). Two of the 11 genes were annotated as Waldenstrom's macroglobulinemia susceptibility and other two were annotated as psoriasis susceptibility. In addition, 11 genes that have been classified as hypothetical proteins (Categories V and VI) in H-InvDB 6.2 turned out to be noncoding RNA candidates (Supplementary Table S2). Four of these genes were annotated as similar to functional noncoding RNAs.

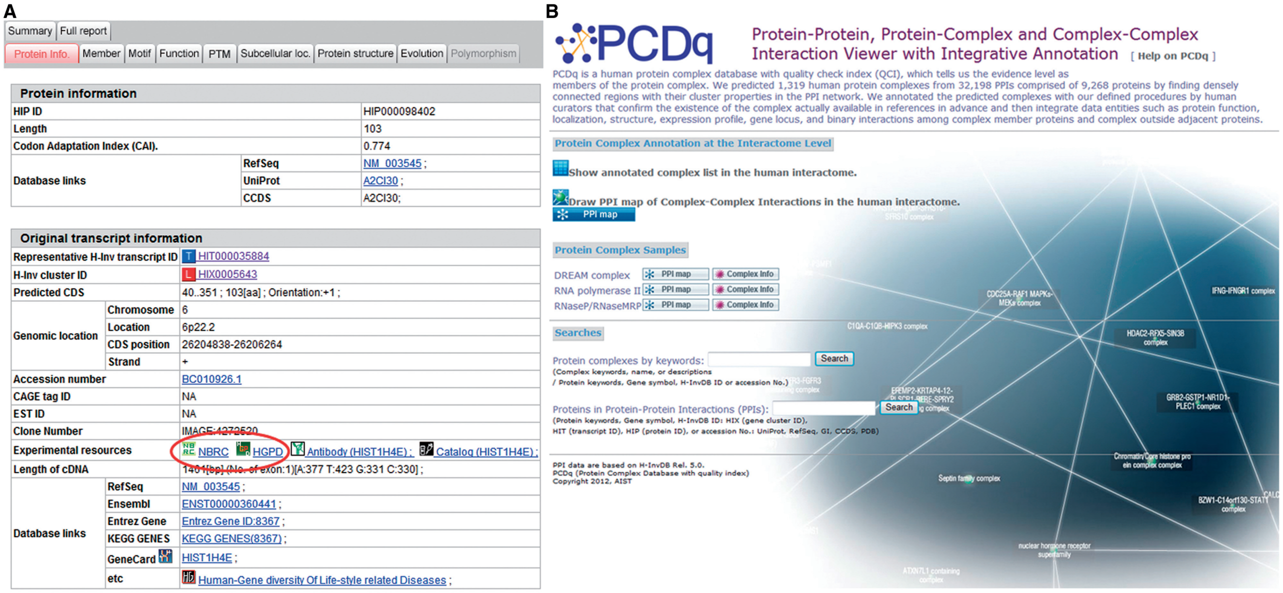
New features

We had originally developed annotation viewers for transcriptomes and genomes, called 'Transcript view' and 'Locus view', respectively. In addition to these viewers, we provided a new viewer named 'Protein view' for the annotations of the human proteome (Figure 3A). In Protein view, annotation information of H-Inv proteins (HIPs) is provided. Furthermore, through the web service APIs, a link to GlycoProtein DataBase (27) is added and the glycosylation site is illustrated in the figure of Protein view. As human cDNA clones are necessary for protein expression experiments, we added links to the human cDNA clone databases such as Biological Resource Center (NBRC) and Human Gene and Protein Database (HGPD) (28), which are connected by HMS.

Table 3. Protein category-upgraded genes relating with Mendelian disorders in only H-InvDB 8.0

Category[a]	Number of category-upgraded genes
Upgrade from V or VI to I	11
Upgrade from V or VI to II	209
Upgrade from V or VI to III	13

[a] Definition of category is shown in Table 2.



Protein information

HIP ID		HIP000098402
Length		103
Codon Adaptation Index (CAI).		0.774
Database links	RefSeq	NM_003545 ;
	UniProt	A2CI30 ;
	CCDS	A2CI30;

Original transcript information

Representative H-Inv transcript ID		T HIT000035884
H-Inv cluster ID		L HIX0005643
Predicted CDS		40..351 ; 103[aa] ; Orientation:+1 ;
Genomic location	Chromosome	6
	Location	6p22.2
	CDS position	26204838-26206264
	Strand	+
Accession number		BC010926.1
CAGE tag ID		NA
EST ID		NA
Clone Number		IMAGE:4272520
Experimental resources		NBRC HGPD Antibody (HIST1H4E) ; Catalog (HIST1H4E) ;
Length of cDNA		1461[bp] (No. of exon:1)[A:377 T:423 G:331 C:330] ;
Database links	RefSeq	NM_003545 ;
	Ensembl	ENST00000360441 ;
	Entrez Gene	Entrez Gene ID:8367 ;
	KEGG GENES	KEGG GENES(8367) ;
	GeneCard	HIST1H4E ;
	etc	Human-Gene diversity Of Life-style related Diseases ;

Protein-Protein, Protein-Complex and Complex-Complex Interaction Viewer with Integrative Annotation [Help on PCDq]

PCDq is a human protein complex database with quality check index (QCI), which tells us the evidence level as members of the protein complex. We predicted 1,319 human protein complexes from 32,198 PPIs comprised of 9,268 proteins by finding densely connected regions with their cluster properties in the PPI network. We annotated the predicted complexes with our defined procedures by human curators that confirm the existence of the complex actually available in references in advance and then integrate data entities such as protein function, localization, structure, expression profile, gene locus, and binary interactions among complex member proteins and complex outside adjacent proteins.

Protein Complex Annotation at the Interactome Level

Show annotated complex list in the human interactome.

Draw PPI map of Complex-Complex Interactions in the human interactome.

PPI map

Protein Complex Samples

DREAM complex PPI map Complex Info
RNA polymerase II PPI map Complex Info
RNaseP/RNaseMRP PPI map Complex Info

Searches

Protein complexes by keywords: Search
(Complex keywords, name, or descriptions / Protein keywords, Gene symbol, H-InvDB ID or accession No.)

Proteins in Protein-Protein Interactions (PPIs): Search
(Protein keywords, Gene symbol, H-InvDB ID: HIX (gene cluster ID), HIT (transcript ID), HIP (protein ID), or accession No.: UniProt, RefSeq, GI, CCDS, PDB)

PPI data are based on H-InvDB Rel. 5.0.
PCDq (Protein Complex Database with quality index)
Copyright 2012, AIST

SFRS10 complex
CDC25A-RAF1 MAPKs-MEKs complex
IFNG-IFNGR1 complex
C1QA-C1QB-HIPK3 complex
HDAC2-RFX5-SIN3B complex
EFEMP2-KRTAP4-12-PLSCR1-RERE-SPRY2 complex
GRB2-GSTP1-NR1D1-PLEC1 complex
Chromatin/Core histone protein complex complex
Septin family complex
BZW1-C14orf130-STAT1 complex
nuclear hormone receptor superfamily
ATXN7L1 containing complex

Figure 3. Screenshot of a part of protein view and the top page of PCDq. **(A)** Hyperlinks to NBRC and HGPD are shown in a red circle. **(B)** Entrance to PCDq is http://www.h-invitational.jp/hinv/pcdq/.

Using these links, users can access the databases and obtain actual human cDNA clones for various experiments. A new subdatabase was also constructed and connected with H-InvDB. This new subdatabase called Protein Complex Database with quality index (PCDq) (29) is a human protein complex database with complex quality index, which describes evidence levels as subunits (protein members) of the protein complex. From the human PPI network dataset integrated from the six PPI data, human protein complexes were predicted and curated with the literature. Thus, PCDq consists of both known and predicted complexes/subunits (Figure 3B). PCDq is expected to enable users to investigate protein interactions in more detail by protein subunit rather than whole protein.

FUTURE PERSPECTIVES

At present, the identification of all human proteins is proceeding worldwide. H-InvDB will continue to offer tools for proteome studies. For example, we are now collecting information on posttranslational modification. Using feedback from various experimental results at the protein level, we intend to develop H-InvDB as the best central hub for human omics study. In addition, personal genome annotation such as the prediction of disease susceptibility using individual gene mutations will be much required. Therefore, we intend to expand the field of personal genomics in future. In addition to the web service APIs of the present H-InvDB, we will provide annotation data in the Resource Description Framework (RDF) (http://www.w3.org/RDF/). We aim to improve the efficiency of accessing molecular biological data by integrating international databases in a more sophisticated manner using this semantic web technology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2.

ACKNOWLEDGEMENTS

The authors acknowledge all the members of the H-Invitational consortium and the Genome Information Integration Project (GIIP) for participating in the annotation work of human full-length cDNAs. The authors also acknowledge all the staffs of H-InvDB, especially Tomohiro Endo and Kentarou Mamiya for their technical supports.

FUNDING

National Institute of Advanced Industrial Science and Technology (AIST) and the life science database project of the Ministry of Economy, Trade, and Industry (METI) of Japan. Funding for open access charge: AIST.

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Ota,T., Suzuki,Y., Nishikawa,T., Otsuki,T., Sugiyama,T., Irie,R., Wakamatsu,A., Hayashi,K., Sato,H., Nagai,K. *et al.* (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.*, **36**, 40–45.
3. Kikuno,R., Nagase,T., Nakayama,M., Koga,H., Okazaki,N., Nakajima,D. and Ohara,O. (2004) HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEppi and ROUGE. *Nucleic Acids Res.*, **32**, D502–D504.

4. Temple,G., Gerhard,D.S., Rasooly,R., Feingold,E.A., Good,P.J., Robinson,C., Mandich,A., Derge,J.G., Lewis,J., Shoaf,D. *et al.* (2009) The completion of the Mammalian Gene Collection (MGC). *Genome Res.*, **19**, 2324–2333.
5. Mewes,H.W., Dietmann,S., Frishman,D., Gregory,R., Mannhaupt,G., Mayer,K.F., Munsterkotter,M., Ruepp,A., Spannagl,M., Stumpflen,V. *et al.* (2008) MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res.*, **36**, D196–D201.
6. Zhang,Q.H., Ye,M., Wu,X.Y., Ren,S.X., Zhao,M., Zhao,C.J., Fu,G., Shen,Y., Fan,H.Y., Lu,G. *et al.* (2000) Cloning and functional analysis of cDNAs with open reading frames for 300 previously undefined genes expressed in CD34+ hematopoietic stem/progenitor cells. *Genome Res.*, **10**, 1546–1560.
7. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
8. Yamasaki,C., Murakami,K., Takeda,J., Sato,Y., Noda,A., Sakate,R., Habara,T., Nakaoka,H., Todokoro,F., Matsuya,A. *et al.* (2010) H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic Acids Res.*, **38**, D626–D632.
9. Yamasaki,C., Kawashima,H., Todokoro,F., Imamizu,Y., Ogawa,M., Tanino,M., Itoh,T., Gojobori,T. and Imanishi,T. (2006) TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB. *Nucleic Acids Res.*, **34**, W345–W349.
10. Tanino,M., Debily,M.A., Tamura,T., Hishiki,T., Ogasawara,O., Murakawa,K., Kawamoto,S., Itoh,K., Watanabe,S., de Souza,S.J. *et al.* (2005) The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. *Nucleic Acids Res.*, **33**, D567–D572.
11. Matsuya,A., Sakate,R., Kawahara,Y., Koyanagi,K.O., Sato,Y., Fujii,Y., Yamasaki,C., Habara,T., Nakaoka,H., Todokoro,F. *et al.* (2008) Evola: ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.*, **36**, D787–D792.
12. Shimada,M.K., Matsumoto,R., Hayakawa,Y., Sanbonmatsu,R., Gough,C., Yamaguchi-Kabata,Y., Yamasaki,C., Imanishi,T. and Gojobori,T. (2009) VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. *Nucleic Acids Res.*, **37**, D810–D815.
13. Takeda,J., Suzuki,Y., Sakate,R., Sato,Y., Gojobori,T., Imanishi,T. and Sugano,S. (2010) H-DBAS: human-transcriptome database for alternative splicing: update 2010. *Nucleic Acids Res.*, **38**, D86–D90.
14. Imanishi,T. and Nakaoka,H. (2009) Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases. *Nucleic Acids Res.*, **37**, W17–W22.
15. Kodama,Y., Mashima,J., Kaminuma,E., Gojobori,T., Ogasawara,O., Takagi,T., Okubo,K. and Nakamura,Y. (2012) The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. *Nucleic Acids Res.*, **40**, D38–D42.
16. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
17. The UniProt Consortium. (2011) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
18. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2011) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
19. Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
20. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
21. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
22. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
23. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
24. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
25. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
26. Amberger,J., Bocchini,C. and Hamosh,A. (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). *Hum. Mutat.*, **32**, 564–567.
27. Kaji,H., Saito,H., Yamauchi,Y., Shinkawa,T., Taoka,M., Hirabayashi,J., Kasai,K., Takahashi,N. and Isobe,T. (2003) Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat. Biotechnol.*, **21**, 667–672.
28. Maruyama,Y., Kawamura,Y., Nishikawa,T., Isogai,T., Nomura,N. and Goshima,N. (2012) HGPD: Human Gene and Protein Database, 2012 update. *Nucleic Acids Res.*, **40**, D924–D929.
29. Kikugawa,S., Nishikata,K., Murakami,K., Sato,Y., Suzuki,M., Altaf-Ul-Amin,M., Kanaya,S. and Imanishi,T. (2012) PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-Invitational protein-protein interactions integrative dataset. *BMC Syst. Biol.*, in press.